在 Deepfake 检测中保持公平性泛化

摘要

尽管近年来已经开发出了有效的深度伪造检测模型，但最近的研究表明，这些模型可能会导致种族和性别等人口统计群体之间的不公平表现差异。这可能会导致特定群体面临不公平的目标或被排除在检测之外，可能会允许错误分类的深度伪造操纵公众舆论并破坏对模型的信任。解决这个问题的现有方法是提供公平损失函数。它显示了良好的公平性能域内评估，但不保持公平跨域测试。这凸显了公平泛化在打击深度伪造方面的重要性。在这项工作中，我们提出了第一种方法，通过同时考虑特征、损失和优化方面来解决深度伪造检测中的公平性泛化问题。我们的方法采用解纠缠学习来提取人口统计和领域不可知的伪造特征，将它们融合在一起，以鼓励在平坦的损失环境中进行公平学习。在突出的深度伪造数据集上进行的广泛实验证明了我们方法的有效性，在跨域深度伪造检测期间保持公平性方面超过了最先进的方法。

关键词： 深度伪造检测；公平性学习；解耦学习

1 引言

Deepfakes 是“深度学习（Deeplearning）”和“假（Fake）”的合成词，已经成为当代技术中一个火热而又令人担忧的方面。这些是人工智能生成或操纵的媒体（例如，图像，视频）通过深度神经网络（例如，变分自动编码器 [41] ，生成对抗网络 [25] ，扩散模型 [5] ）生成，这些模型的效果看起来非常真实，通常使得被操作的个人参与他们从未参与的行动或说出他们从未说过的话。虽然 deepfakes 为创意内容和娱乐打开了大门，但恶意使用 deepfakes 可能会导致错误信息，隐私泄露，甚至政治操纵，减少信任并产生混乱 [46]， [33] 。为了对抗欺骗性的深度伪造的传播，出现了一个新兴的深度伪造检测方法领域，这些方法是数据驱动和基于深度学习的 [2,3,7,8,10,12,15,23,27–29,31,36,44,49–53]。然而，最近的研究和报告 [39] [32] 揭示了当前深度伪造检测方法中的公平性问题。一个重要的问题是，在评估不同的人口统计群体（包括性别、年龄和种族）时，表现不一致。例如，一些最先进的检测器在评估深度伪造时表现出更高的准确性，其特征是与肤色较深的人相比，肤色较浅的人 [39] [14]。这使得攻击者能够针对特定人群生成有害的深度伪造，以逃避检测。Ju [24] 等人提出了一种解决深度伪造检测中公平性的初始算法级方法。他们表明，所提出的 DAW-FDD 模型可以在域内（训练和测试数据由相同的伪造技术生成）评估场景下表现出最佳的公平性性能。然而，在实践中，我们发现当测试由未知的伪造技术生成的数据时他们的方法不能保持跨域评估的公平性。因此，实现公平泛化至关重要，如果没有这样的推广，当前公平的深度伪造检测方法很容易过时。

2 相关工作

2.1 深度伪造检测

现有的深度伪造检测方法中最大的一部分属于数据驱动的类别，包括 [31] [36] 。这些方法利用在真实和深度伪造视频上训练的各种类型的深度神经网络（DNN）来捕获特定的可辨别的伪影。虽然这些方法在域内评估方面取得了令人满意的性能，但它们在跨域测试方面的性能急剧下降。为了解决泛化问题，解纠缠学习 [45] 被广泛用于通过提取相关特征同时消除不相关特征来进行伪造检测。例如，Hu [15] 引入了一个解纠缠框架来自动定位与伪造相关的区域，Zhang 等人 [51] 通过辅助监督增强了泛化能力。Liang 等人 [27] 提出了一个框架，通过内容一致性和全局表示对比约束来提高特征独立性。Yan 等人 [49] 通过专门利用与伪造相关特征分离的常见伪造特征来扩展此框架。

2.2 深度伪造的公平性检测

最近的研究提到了深度伪造检测中的公平性问题 [32]。Trinh 等人 [39] 确定了深度伪造数据集和检测模型中的偏差，揭示了各亚组之间的显著错误率差异。Hazirbas 等人的研究中报告了类似的观察结果。Pu 等人 [35] 评估了 MesoInception-4 深度伪造检测模型在 FF++ 上的公平性，发现它对男女都不公平。Xu 等人 [48] 对深度伪造检测中的偏差进行了全面分析，丰富了具有不同注释的数据集，以支持未来的研究。此外，Nadimpalli 等人 [34] 强调了数据集和检测模型中的实质性偏倚，引入了性别平衡的数据集以减轻基于性别的性能偏倚。然而，这种方法只产生了适度的改进，并需要大量的数据注释。Ju 等人 [24] 专注于增强同一数据域内的公平性，但没有解决跨域测试中的公平性，这是我们论文的中心焦点。

3 本文方法

3.1 本文方法概述

本文提出的模型整体架构图如 1所示，包括三个模块：解纠缠学习，公平学习和优化。解纠缠学习模块的目的是从输入图像中提取域不可知的伪造和人口统计特征。公平学习模块利用这两种类型的特征来开发公平分类器。两个学习模块都由优化模块监督，从而在模型训练期间增强公平性泛化。本文将在下面的部分深入研究每个模块的细节。

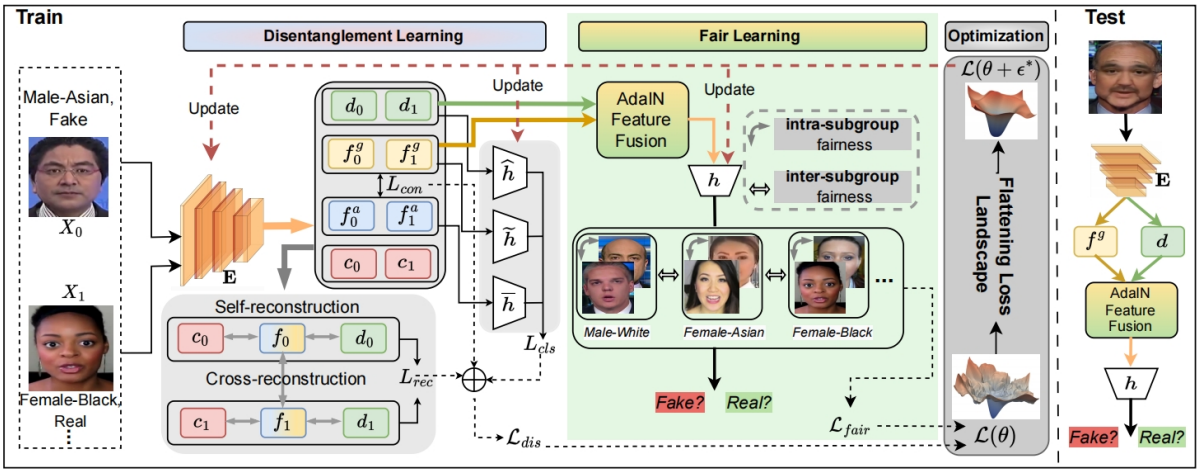

图 1. 模型整体架构图

3.2 Exposing Demographic & Forgery Features

该架构提出了一个解纠缠学习模块来提取人口统计特征（公平性）和域不可知伪造特征（泛化），为了实现这点，使用成对的图像 $(X_i, X_{i'})$，其中，$i, i' \in \{1, \cdots, n\}, i \neq i'$ 并且 X_i 和 $X_{i'}$ 互为真假，即当 X_i 为真的时候，$X_{i'}$ 为假；当 X_i 为假的时候，$X_{i'}$ 为真。每个图像由 Encoder$E(\bullet)$ 处理，Encoderr$E(\bullet)$ 包含 3 个不同的编码器，每个编码器分别负责提取内容特征 c（与图像背景相关）、伪造特征 d 和人口统计特征。伪造特征又分为特定域伪造特征 f_i^a 和域不可知伪造特征 f_i^g，其公式如下：

$$c_i, f_i^a, f_i^g, d_i = E\left(X_i\right)$$

3.3 分类损失（Classification Loss）

对于解耦特定域的伪造、域不可知的伪造和人口统计特征通常设计到对每个特征使用交叉熵损失（cross-entropy loss）。然而，深度伪造数据集通常都有人口统计子组分布不平衡的现象，这是实现检测公平性的一个基本问题。此外，传统的交叉熵损失训练往往会导致大多数子组的样本进行过拟合，从而导致其不适合学习公平的人口统计特征的表示。为了解决这些挑战，我们受 [1] 的启发，采取了人口分布感知边际损失，如下公式所示：

$$M\left(\hat{h}\left(d_i\right), D_i\right) = -\log \frac{e^{\hat{h}^{D_{i(d_i)} - D_i}}}{e^{\hat{h}^{D_{i(d_i)} - D_i}} + \sum_{p \neq D_i} e^{\hat{h}^{p(d_i)}}}$$

其中，$p = \frac{\delta}{n_p^{1/4}}$ 是的人口统计子组相关边际，δ 是常数，n_p 表示子组训练数据点的个数，$\hat{h}$ 是 d_i 的分类头，$\hat{h}^p$ 表示 p 的输出。

通过合并这种边际损失，我们利用较大的边际 p 来改善具有较小 n_p 的的少数子组的泛化，从而促进无偏差的人口统计特征表示。因此，总分类损失为：

$$L_{cls} = C\left(\widetilde{h}\left(f_i^g\right), Y_i\right) + \rho_1 C\left(\bar{h}\left(f_i^a\right), A_i\right) + \rho_2 M\left(\hat{h}\left(d_i\right), D_i\right)$$

其中，$C(\bullet,\bullet)$ 为交叉熵损失，$\sim h$ 和 $\widetilde{h}$ 分别是两个支持向量机的分类头，ρ_1 和 ρ_2 是两个权衡超参数（Tade-off Hyperparameters）。用上述的分类损失进行训练，可以使得编码器获得特定的特征信息，增强模型的泛化能力。

3.4 对比损失（contrast Loss）

考虑到分类损失侧重于单个图像，忽略了在增强编码器表示能力方面起关键作用的图像相关性。受对比学习的启发 [49] [40]，我们可以引入对比损失（Contrast Loss）来解决这一问题，其计算公式如下:

$$L_{con} = [b + ||f_{anchor} - f_{+}||_2 - ||f_{anchor} - f_{-}||_2]_{+}$$

其中，f_{anchor} 代表图像的锚伪造特征，f_{+} 和 f_{-} 分别代表来自同一来源的正面对应和来自不同来源的负面对应。是一个超参数，$[\bullet]_{+} = max\{0, \bullet\}$ 是一个铰函数。在实验中，我们对特定域和不可知域的伪造特征都使用了 L_{con}。对于特定于域的伪造特征，源被认为是伪造域，对比损失激励编码器学习特定的伪造表示。对于与领域无关的伪造特征，源可以是真实的，也可以是虚假的，而损失则鼓励编码器学习一种不与任何特定伪造方法相关联的可推广的表示。

3.5 重建损失（Reconstruction Loss）

为了保持提取特征的完整性，并在像素级上保持原始图像和重建图像之间的一致性，我们采用了重建损失。其公式为:

$$L_{rec} = ||X_i - D(c_i, f_i, d_i)||_1 + ||X_i - D(c_i, f_{i'}, d_i)||_1$$

其中，$D(\bullet,\bullet,\bullet)$ 是负责使用解纠缠特征表示重建图像的解码器。在 L_{rec} 损失中，第一项是自重构损失（Self-reconstruction Loss），它利用输入图像的潜在特征最小化重构误差。第二项是交叉重建损失（Cross-reconstruction Loss），它通过结合合作伙伴的伪造特征来惩罚重建错误。这两个损失一起工作来改善特征的解纠缠。

3.6 解耦损失（Disentanglement Loss）

因此，解决人口统计特征和伪造特征的解耦损失如以下公式所示：

$$L_{dis} = \frac{1}{n}\sum_{i}[L_{cls} + \rho_3 L_{con} + \rho_4 L_{rec}]$$

其中，ρ_3 和 ρ_4 是权衡超参数（Tade-off Hyperparameters）

3.7 Fair Learning under Generalization

我们获得了与领域无关的伪造特征和人口统计特征后，使用自适应实例归一化 (AdaIN) [22] 将它们结合起来，以达到公平学习的目的。融合的特征 I_i 可以形成如下，

$$I_i = \sigma(d_i)\left(\frac{(f_i^g - \mu(f_i^g))}{\sigma(f_i^g)}\right) + \mu(d_i)$$

其中，$\mu(\bullet)$ 和 $\sigma(\bullet)$ 分别为每个通道独立计算输入特征跨空间维度的均值和标准差。由于大部分深度伪造生成方法通常会修改图像的面部区域，其中就包含用于缺点的人口统计信息的基本特征。

3.8 公平性损失（Fairness Loss）

传统的实现公平学习方法，如 [42] [43]，通常涉及在学习目标中添加公平惩罚。然而，这些方法只能确保特定公平度量的公平性，如人口均等 [6] 或均等赔率 [13]，这限制了模型的公平性可扩展性及其处理新数据集的能力。此外，即使整个深度假数据集已经平衡了 Fake 的和 Real 的例子，不平衡仍然可能存在于人口统计子组中，这可能导致这些子组内的偏见学习。为了解决这些问题，受到 [24] [16–21] 的启发，我们引入了双层次公平损失（Bi-level Fairness loss），其计算公式如下所示：

$$L_{fair} = \min_{\eta \in \mathbb{R}} \eta + \frac{1}{\alpha \ |J|} \sum_{i=1}^{|J|} [L_i - \eta]_+$$

$$s.t. \ L_j = \min_{\eta_j \in \mathbb{R}} \eta_j + \frac{1}{\alpha' \ |J_j|} \sum_{i \ : \ D_i = \ J_j} [C\left(h\left(I_i\right), \ Y_i\right) - \ \eta_j]_+$$

其中，$|J|$ 表示集合的大小，每个子组 $J_j \in J$ 并且 $|J_j|$ 表示 J_j 中训练样本的个数。h 是 I_i 的分类头，与其他头共享相同的 MLP 架构，$\alpha, \alpha' \in (0, 1)$ 是两个超参数。外层次公式 L_{fair} 的灵感来自于公平性风险方法 [47]，旨在促进子群体间的公平性。内层公式 $s.t. \ L_j$ 受到分布鲁棒优化（即条件风险值 [21]）的启发，该优化增强了子组内 Real 和 Fake 样本的真实性，从而增强了模型的鲁棒性。

3.9 联合优化（Joint Optimization）

对上述两个模块进行联合优化。为了避免图 2 中描述的大量尖锐和狭窄的最小值，我们利用锐度感知最小化方法 [9] 来平坦化损失景观（loss landscape）。具体来说，将整个框架的模型权重表示为，通过确定扰动 $\quad$ 的最佳来最大化损失来实现平坦化，定义为：

$$\epsilon^* \ = \ arg \max_{||\varepsilon||_2 \leq \ \gamma} \underbrace{(L_{dis} + \ \lambda L_{fair})}_{L} (\theta \ + \ \varepsilon)$$

$$\approx arg \max_{||\varepsilon||_2 \leq \ \gamma} \epsilon^T \nabla_\theta L = \ \gamma \ sign\left(\nabla_\theta L\right)$$

其中，γ 是控制扰动幅度的超参数，λ 是权衡超参数。近似是用一个一阶泰勒展开得到的，假设 $\nabla_\theta L$ 是相对于 L 的梯度。因此，通过解决以下问题来更新模型权重：

$$\min_\theta L\left(\theta \ + \ \epsilon^*\right)$$

因此，沿梯度范数方向的扰动会显著增加损失值，然后使模型在公平性方面更具泛化性。

4 复现细节

4.1 与已有开源代码对比

本文复现了作者所提出的模型架构，并且改用了 Resnet-50 网络替代原有的 Xception 主干网络。同时为了增强模型对新型伪造技术的适应性和鲁棒性，在原有的数据集中利用最近比较新的 talkinghead 方法，例如，SadTalk、StyleSync、Emotalk 等方法替换掉了 20% 的原始图片。

4.2 实验环境搭建

所有实验均基于 PyTorch，使用 Tesla P100-PCIE-16GB 进行训练。对于训练，我们固定的批大小为 8，epoch 为 50，使用学习率为 $\beta = 5 \times 10^{-4}$ 的 SGD 优化器。对于整体损失，我们设 $\lambda = 1$，$\gamma = 0.05$(平坦化损失中扰动的邻域大小)，L_{cls} 中的 $\rho_1 = 0.1$ 和 $\rho_2 = 0.1$，L_{dis} 中的 $\rho_3 = 0.05$ 和 $\rho_4 = 0.3$，L_{con} 中的 $b = 3.0$，$M\left(\hat{h}\left(d_i\right), D_i\right)$ 中的 $\delta = 2.89$（根据人口统计学样本分布)。α 和 α' 在 $0, 1, 0.3, 0.5, 0.7, 0.9$ 进行调谐。在 [24] 之后，最终的 α 和 α' 是根据一个预设规则确定的，该规则允许从相应的' DAW-FDD ' 方法中验证集中的总体 AUC 降低 5%，同时最小化 F_{FPR} 上的交集组。

4.3 数据集设置

为了验证我们提出的方法的公平性泛化能力，我们在最广泛使用的基准测试 FaceForensics++(FF++) [37] 训练我们的模型，并在 FF++ +、DeepfakeDetection (DFD) [11]、Deepfake detection Challenge (DFDC) [4] 和 Celeb-DF [26] 上进行测试。由于原始数据集不具有每个视频或图像的人口统计信息，因此我们遵循 Ju [24] 等人的进行数据处理、数据注释和敏感属性组合 (Intersection)。因此，子集组包含亚洲人男性 Male-Asian (M-A)、白人男性 Male-White (M-W)、黑人男性 Male-Black(M-B)、其他种族男性 Male-Others (M-O)、亚洲人女性 Female-Asian (F-A)、白人女性 Female-White (F-W)、黑人女性 Female-Black (F-B) 和其他种族女性 Female-Others (F-O)。

4.4 评估指标

为了进行检测比较，我们使用曲线下面积 (Area under Curve, AUC) 来对我们的方法与之前的工作进行基准测试，这与之前的工作采用的检测评估方法一致 [49] [30]。关于公平性，我们使用四个不同的公平性指标来评估我们提出的方法的有效性。具体来说，我们报告了相等假阳性率 (F_{FPR}) [24]，最大相等赔率 (F_{MED}) [42]，人口平价 (F_{DP}) [43][37] 和总体准确性平等 (F_{OAE}) [42]。这些公平性指标计算公式如下所示：

$$F_{\text{FPR}} = \sum_{J_j \in \mathcal{J}} \left| \frac{\sum_{j=1}^{n} I\left[\hat{Y}_j = 1, D_j = J_j, Y_j = 0\right]}{\sum_{j=1}^{n} I\left[D_j = J_j, Y_j = 0\right]} - \frac{\sum_{j=1}^{n} I\left[\hat{Y}_j = 1, Y_j = 0\right]}{\sum_{j=1}^{n} I\left[Y_j = 0\right]} \right|$$

$$F_{\text{OAE}} = \max_{J_j \in \mathcal{J}} \left\{ \frac{\sum_{j=1}^{n} I\left[\hat{Y}_j = Y_j, D_j = J_j\right]}{\sum_{j=1}^{n} I\left[D_j = J_j\right]} - \min_{J'_j \in \mathcal{J}} \frac{\sum_{j=1}^{n} I\left[\hat{Y}_j = Y_j, D_j = J'_j\right]}{\sum_{j=1}^{n} I\left[D_j = J'_j\right]} \right\}$$

$$F_{\text{DP}} = \max_{k \in \{0,1\}} \left\{ \max_{J_j \in \mathcal{J}} \frac{\sum_{j=1}^{n} I\left[\hat{Y}_j = k, D_j = J_j\right]}{\sum_{j=1}^{n} I\left[D_j = J_j\right]} - \min_{J'_j \in \mathcal{J}} \frac{\sum_{j=1}^{n} I\left[\hat{Y}_j = k, D_j = J'_j\right]}{\sum_{j=1}^{n} I\left[D_j = J'_j\right]} \right\}$$

$$F_{\text{MED}} = \max_{k,k' \in \{0,1\}} \left\{ \max_{J_j \in \mathcal{J}} \frac{\sum_{j=1}^{n} I\left[\hat{Y}_j = k, Y_j = k', D_j = J_j\right]}{\sum_{j=1}^{n} I\left[D_j = J_j, Y_j = k'\right]} - \min_{J'_j \in \mathcal{J}} \frac{\sum_{j=1}^{n} I\left[\hat{Y}_j = k, Y_j = k', D_j = J'_j\right]}{\sum_{j=1}^{n} I\left[D_j = J'_j, Y_j = k'\right]} \right\}$$

其中，D 是人口统计变量，$\mathcal{J}$ 是子组的集合，每个子组 $J_j \in \mathcal{J}$。FFPR 衡量不同群体之间相对于整体人群的假阳性率（FPR）差异。F_{OAE} 衡量所有人口统计群体中的最大准确性差

距。F_{DP} 衡量所有人口统计群体中预测率的最大差异。F_{MED} 捕捉在比较不同人口统计群体时预测结果（无论是正向还是负向）中的最大差异。

4.5 创新点

本文通过复现并改进原有模型架构，采用 ResNet-50 网络替代 Xception 主干网络，以及在数据集中融入最新的 talkinghead 方法如 SadTalk、StyleSync、Emotalk 等，替换 20% 的原始图片，以此增强模型对新型伪造技术的适应性和鲁棒性，从而提升了模型在深度伪造检测领域的泛化能力和实用性。

5 实验结果分析

本部分对实验所得结果进行分析，详细对实验内容进行说明，实验结果进行描述并分析。实验训练的损失图如图 2所示，有图可知，随着训练次数的增加，在 30 个 epoch 之后逐渐收敛。公平性泛化性能对比，以 Xception 网络为例，从表 1 中可以看出，我们复现的方法相对于 DAW-FDD [49] 的方法具有更强的公平性泛化能力，同时也获得了最好的 AUC 结果。具体来说，我们的方法在 DFDC 上的 FPR 提升了 32.1%，在 DFD 上的 FPR 提升了 6.4%。此外，该方法在 Celeb-DF 上的公平性性能均不如 DAW-FDD 方法，我们分析可能是因为在训练的时候没有完全训练原我们所提到的 100 个 epoch 导致训练精度不如原方法。

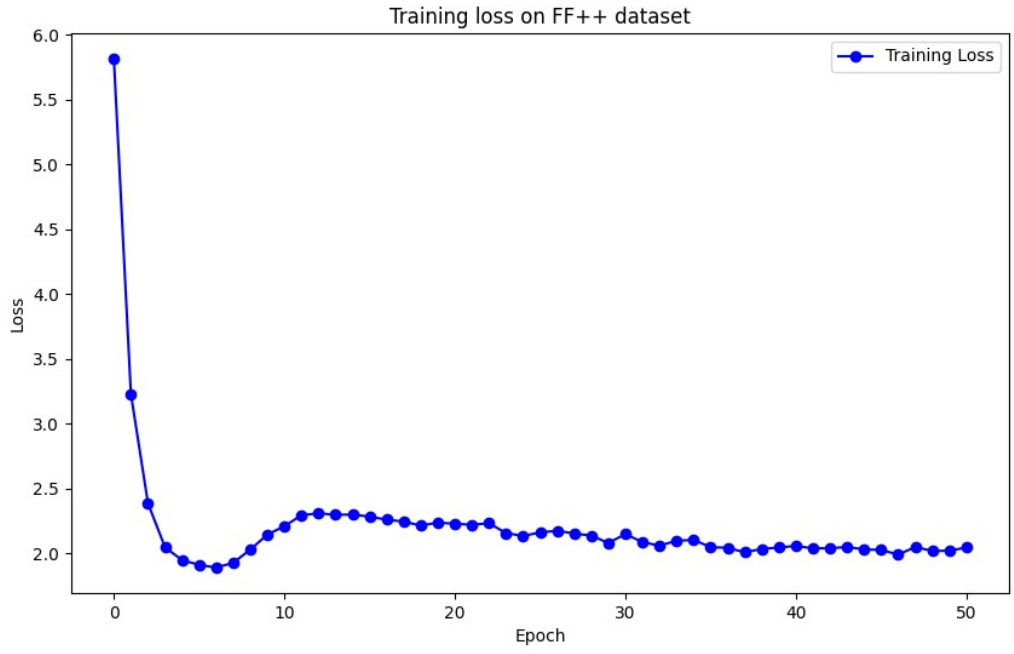

图 2. 模型在各数据集子集上的 AUC 性能

我们针对不同的数据集的属性子集进行测试，其 AUC 结果如图 3所示，从图中可知，我们的方法，在 FF++、DFD、Celeb-DF 数据集上的效果均超越方法。

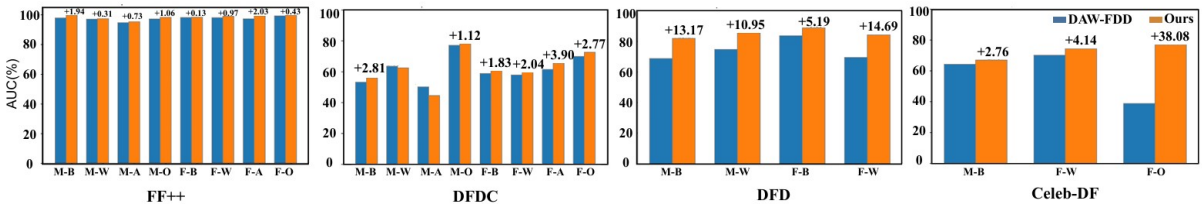

图 3. 模型在各数据集子集上的 AUC 性能

公平性泛化性能对比，以 Xception 网络为例，从表 1 中可以看出，我们复现的方法相对于 DAW-FDDD [49] 的方法具有更强的公平性泛化能力，同时也获得了最好的 AUC 结果。具体来说，我们的方法在 DFDC 上的 FPR 提升了 32.1%，在 DFD 上的 FPR 提升了 6.4%。此外，该方法在 Celeb-DF 上的公平性性能均不如 DAW-FDD 方法，我们分析可能是因为在训练的时候没有完全训练原我们所提到的 100 个 epoch 导致训练精度不如原方法。

表 1. 域内 (FF++) 和跨域 (DFDC、Celeb-DF 和 DFD) 的性能比较

Dataset	**Source**	**FPR**	**MEO**	**DP**	**OAE**	**AUC**
FF++	DAW-FDD	31.31%	17.69%	11.12%	10.08%	92.77%
FF++	Ours	14.61%	14.61%	17.83%	11.62%	98.73%
DFDC	DAW-FDD	52.77%	37.78%	13.87%	30.30%	56.72%
DFDC	Ours	20.67%	25.23%	14.91%	17.74%	59.26%
DFD	DAW-FDD	35.14%	28.52%	15.31%	12.95%	74.34%
DFD	Ours	28.74%	28.74%	14.58%	12.31%	82.88%
Celeb-DF	DAW-FDD	33.69%	33.37%	27.45%	35.42%	62.66%
Celeb-DF	Ours	27.55%	25.65%	17.74%	58.444%	74.42%

损失景观可视化的结果如图 4所示。图 4直观地说明了我们的方法的损失情况，在没有 loss flattening 的情况下，可视化结果的景观清晰，有无数的波峰和波谷。这样的尖锐图形可能会将模型困在次优的极小值中，导致不一致的泛化。然而，在扁平化之后，景观变得平滑，这意味着更容易的优化路径可能会带来更好的训练和泛化。这种可视化强调了联合优化在我们的方法中对增强公平性泛化的重要性。

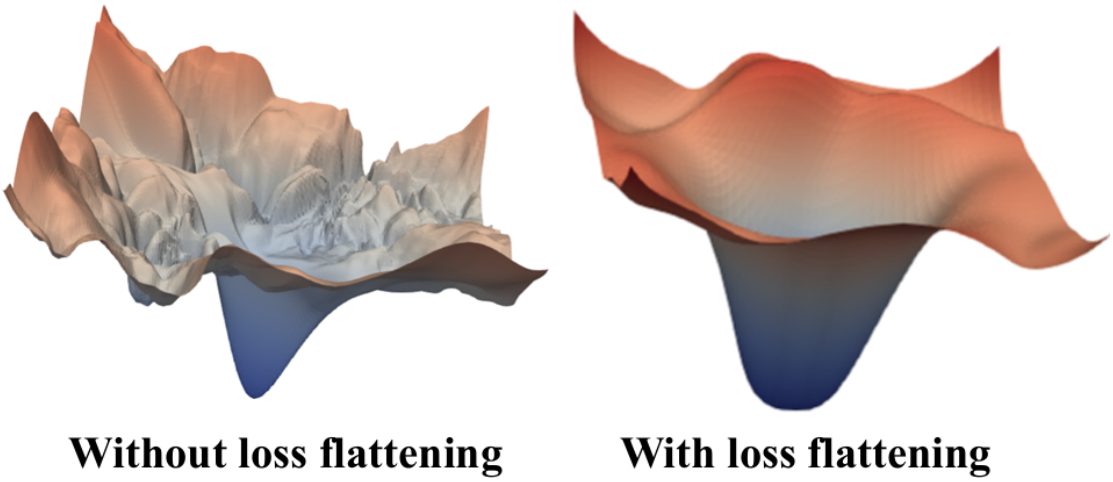

图 4. 损失景观图

为了更直观的证明我们的方法的有效性，我们将 DAW-FDD [49] 我们的方法的 Grad-CAMD [38] 可视化，如图 5所示。Grad-CAM 表明 DAW-FDD 具有公平损失作为约束，在域内表现良好。一旦数据被看不见，它就失去了公平的检测能力。相反，无论数据集如何，我们的方法的激活区域都展示了对面部显著特征的一致模型关注。

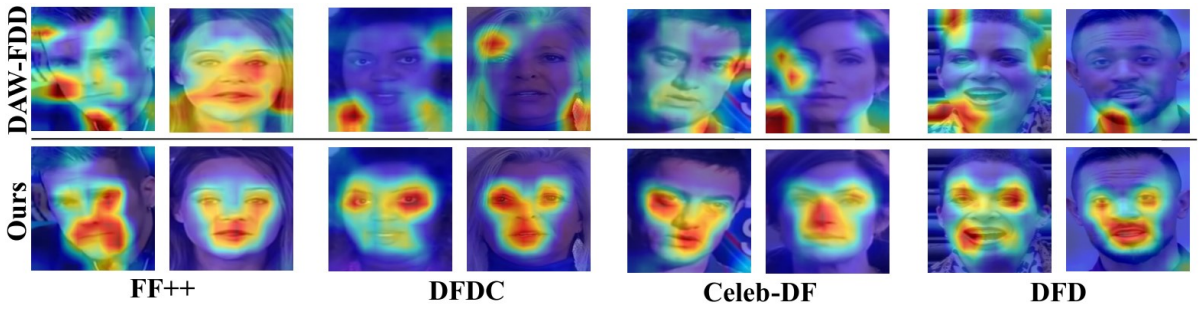

图 5. Grad-CAM 可视化结果

如图 6所示，展示了两种模型方法对特征关注的可视化结果。DAW-FDD 的抽象模式和突出显示的区域 (第二列) 显示了对面部特征的广泛强调，而没有特定的目标。相比之下，我们的解纠缠特征展示了不同的焦点区域: 伪造特征 (第三列) 和人口特征 (最后一列) 主要突出了面部区域，而内容特征 (第四列) 则面向背景。这种差异强调了整合伪造和人口特征以及消除内容特征以促进更公平学习的重要性。

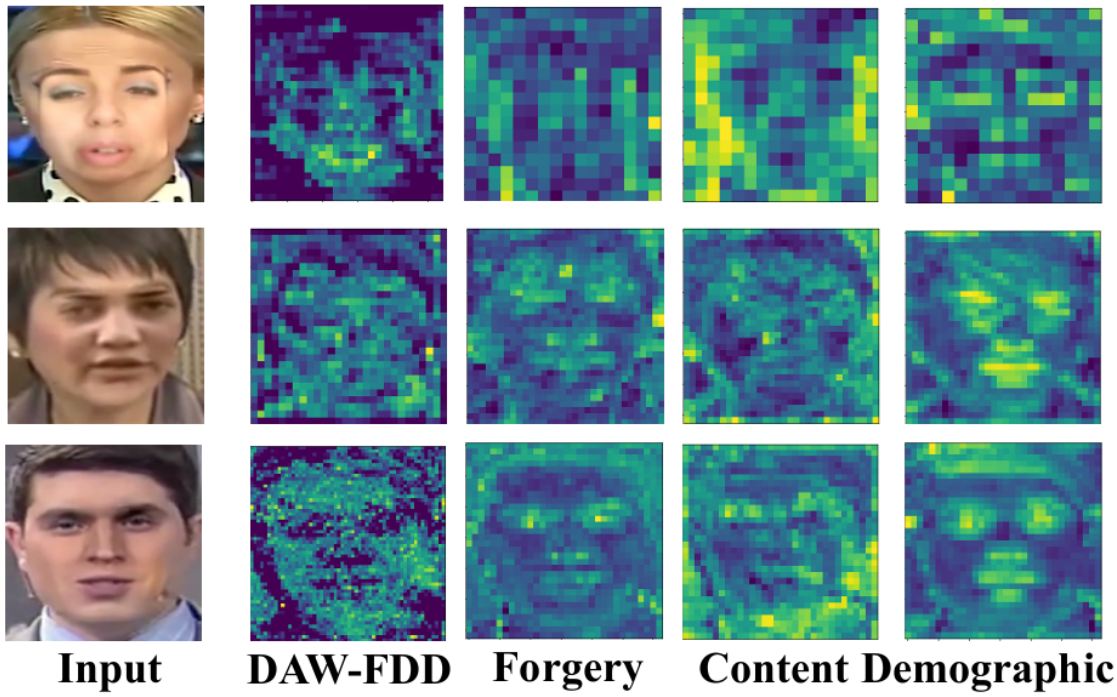

图 6. 特征可视化

6 总结与展望

总的来说，我们复现的方法达到了在大多数公平指标上优于所比较的方法，在公平泛化和 AUC 方面都取得了最好的成绩。虽然目前用于增强深度伪造检测公平性的方法在特定领域内表现良好，但在不同领域进行测试时，它们很难保持公平性。认识到这一限制，我们引入了一个创新的框架，旨在解决深度伪造检测中的公平性泛化挑战。通过结合解纠缠学习和公平学习模块，我们的方法既保证了泛化性，又保证了公平性。此外，我们还结合了损失平坦化策略来简化这些模块的优化过程，从而实现了逆溃公平性泛化。在不同深度伪造数据集上的实验结果显示，我们的方法在不同领域具有卓越的公平性维护能力。

参考文献

[1] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in Neural Information Processing Systems*, 32, 2019.

[2] H. Chen, P. Zheng, X. Wang, S. Hu, B. Zhu, J. Hu, X. Wu, and S. Lyu. Harnessing the power of text-image contrastive models for automatic detection of online misinformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 923–932, 2023.

[3] T. Chen, S. Yang, S. Hu, Z. Fang, Y. Fu, X. Wu, and X. Wang. Masked conditional diffusion model for enhancing deepfake detection. *arXiv preprint arXiv:2402.00541*, 2024.

[4] Deepfake Detection Challenge. Deepfake detection challenge. https://www.kaggle.com/c/deepfake-detection-challenge, 2021. Accessed: 2021-04-24.

[5] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794, 2021.

[6] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

[7] B. Fan, S. Hu, and F. Ding. Synthesizing black-box anti-forensics deepfakes with high visual quality. In *ICASSP*, 2024.

[8] B. Fan, Z. Jiang, S. Hu, and F. Ding. Attacking identity semantics in deepfakes via deep feature fusion. In *2023 IEEE 6th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 114–119, 2023.

[9] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2020.

[10] M. Goebel, L. Nataraj, T. Nanjundaswamy, T. M. Mohammed, S. Chandrasekaran, and B. Manjunath. Detection, attribution and localization of gan generated images. *arXiv preprint arXiv:2007.10466*, 2020.

[11] Google and Jigsaw. Deepfakes dataset by google & jigsaw. https://ai.googleblog.com/2019/09/contributing-data-to-deepfakedetection.html, 2019.

[12] H. Guo, S. Hu, X. Wang, M.-C. Chang, and S. Lyu. Robust attentive deep neural network for detecting gan-generated faces. *IEEE Access*, 10:32574–32583, 2022.

[13] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

[14] C. Hazirbas, J. Bitton, B. Dolhansky, J. Pan, A. Gordo, and C. C. Ferrer. Towards measuring fairness in ai: The casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):324–332, 2021.

[15] J. Hu, S. Wang, and X. Li. Improving the generalization ability of deepfake detection via disentangled representation learning. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3577–3581, 2021.

[16] S. Hu and G. H. Chen. Distributionally robust survival analysis: A novel fairness loss without demographics. In *Machine Learning for Health*, pages 62–87. PMLR, 2022.

[17] S. Hu, L. Ke, X. Wang, and S. Lyu. Tkmi-ap: Adversarial attacks to top-k multi-label learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7649–7657, 2021.

[18] S. Hu, X. Wang, and S. Lyu. Rank-based decomposable losses in machine learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[19] S. Hu, Z. Yang, X. Wang, Y. Ying, and S. Lyu. Outlier robust adversarial training. In *ACML*, 2023.

[20] S. Hu, Y. Ying, S. Lyu, et al. Learning by minimizing the sum of ranked range. *Advances in Neural Information Processing Systems*, 33:21013–21023, 2020.

[21] S. Hu, Y. Ying, X. Wang, and S. Lyu. Sum of ranked range loss for supervised learning. *The Journal of Machine Learning Research*, 23(1):4826–4869, 2022.

[22] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[23] N. Hulzebosch, S. Ibrahimi, and M. Worring. Detecting cnn-generated facial images in real-world scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 642–643, 2020.

[24] Y. Ju, S. Hu, S. Jia, G. H. Chen, and S. Lyu. Improving fairness in deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4655–4665, 2024.

[25] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[26] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu. Celeb-df: A new dataset for deepfake forensics. In *CVPR*, pages 6, 7, 2020.

[27] J. Liang, H. Shi, and W. Deng. Exploring disentangled content information for face forgery detection. In *European Conference on Computer Vision*, pages 128–145, 2022.

[28] L. Lin, N. Gupta, Y. Zhang, H. Ren, C.-H. Liu, F. Ding, X. Wang, X. Li, L. Verdoliva, and S. Hu. Detecting multimedia generated by large ai models: A survey. *arXiv preprint arXiv:2402.00045*, 2024.

[29] Z. Liu, X. Qi, and P. H. Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8060–8069, 2020.

[30] Y. Luo, Y. Zhang, J. Yan, and W. Liu. Generalizing face forgery detection with high-frequency features. In *CVPR*, 2021.

[31] F. Marra, C. Saltori, G. Boato, and L. Verdoliva. Incremental learning for the detection and classification of gan-generated images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2019.

[32] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence*, pages 1–53, 2022.

[33] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence*, 53(4):3974–4026, 2023.

[34] A. V. Nadimpalli and A. Rattani. Gbdf: Gender balanced deepfake dataset towards fair deepfake detection. *arXiv preprint arXiv:2207.10246*, 2022.

[35] M. Pu, M. Y. Kuan, N. T. Lim, C. Y. Chong, and M. K. Lim. Fairness evaluation in deepfake detection models using metamorphic testing. *arXiv preprint arXiv:2203.06825*, 2022.

[36] W. Pu, J. Hu, X. Wang, Y. Li, S. Hu, B. Zhu, R. Song, Q. Song, X. Wu, and S. Lyu. Learning a deep dual-level network for robust deepfake detection. *Pattern Recognition*, 130:108832, 2022.

[37] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019.

[38] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[39] L. Trinh and Y. Liu. An examination of fairness of ai models for deepfake detection. In *IJCAI*, 2021.

[40] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[41] A. Vahdat and J. Kautz. Nvae: A deep hierarchical variational autoencoder. In *Advances in Neural Information Processing Systems*, volume 33, pages 19667–19679, 2020.

[42] H. Wang, L. He, R. Gao, and F. P. Calmon. Aleatoric and epistemic discrimination in classification. In *ICML*, 2023.

[43] J. Wang, X. E. Wang, and Y. Liu. Understanding instance-level impact of fairness constraints. In *International Conference on Machine Learning*, pages 23114–23130. PMLR, 2022.

[44] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8695–8704, 2020.

[45] X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu. Disentangled representation learning. *arXiv preprint arXiv:2211.11695*, 2022.

[46] X. Wang, H. Guo, S. Hu, M.-C. Chang, and S. Lyu. Gan-generated faces detection: A survey and new perspectives. In *ECAI*, 2023.

[47] R. Williamson and A. Menon. Fairness risk measures. In *International Conference on Machine Learning*, pages 6786–6797. PMLR, 2019.

[48] Y. Xu, P. Terhörst, K. Raja, and M. Pedersen. A comprehensive analysis of ai biases in deepfake detection with massively annotated databases. *arXiv preprint arXiv:2208.05845*, 2022.

[49] Z. Yan, Y. Zhang, Y. Fan, and B. Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22412–22423, October 2023.

[50] S. Yang, S. Hu, B. Zhu, Y. Fu, S. Lyu, X. Wu, and X. Wang. Improving cross-dataset deepfake detection with deep information decomposition. *arXiv preprint arXiv:2310.00359*, 2023.

[51] K.-Y. Zhang, T. Yao, J. Zhang, Y. Tai, S. Ding, J. Li, F. Huang, H. Song, and L. Ma. Face anti-spoofing via disentangled representation learning. In *Computer Vision – ECCV 2020: 16th European Conference*, 2020.

[52] L. Zhang, H. Chen, S. Hu, B. Zhu, X. Wu, J. Hu, and X. Wang. X-transfer: A transfer learning-based framework for robust gan-generated fake image detection. *arXiv preprint arXiv:2310.04639*, 2023.

[53] P. Zheng, H. Chen, S. Hu, B. Zhu, J. Hu, C.-S. Lin, X. Wu, S. Lyu, G. Huang, and X. Wang. Few-shot learning for misinformation detection based on contrastive models. *Electronics*, 13(4):799, 2024.